JMLR: Workshop and Conference Proceedings 101:252–267, 2019 ACML 2019

$\mathcal{X}$-Armed Bandits: Optimizing Quantiles, CVaR and Other Risks

Léonard Torossian LEONARD.TOROSSIAN@INRA.FR
Universite de Toulouse, INRA, France and *Institut de Mathematiques de Toulouse, France*

Aurélien Garivier AURELIEN.GARIVIER@ENS-LYON.FR
Univ. Lyon, ENS de Lyon, France

Victor Picheny VICTOR@PROWLER.IO
PROWLER.io, 72 Hills Road, Cambridge, UK

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

We propose and analyze StoROO, an algorithm for risk optimization on stochastic black-box functions derived from StoOO. Motivated by risk-averse decision making fields like agriculture, medicine, biology or finance, we do not focus on the mean payoff but on generic functionals of the return distribution. We provide a generic regret analysis of StoROO and illustrate its applicability with two examples: the optimization of quantiles and CVaR. Inspired by the bandit literature and black-box mean optimizers, StoROO relies on the possibility to construct confidence intervals for the targeted functional based on random-size samples. We detail their construction in the case of quantiles, providing tight bounds based on Kullback-Leibler divergence. We finally present numerical experiments that show a dramatic impact of tight bounds for the optimization of quantiles and CVaR.

Keywords: Optimistic optimization; Risk-averse solutions; Quantile optimization; CVaR optimization

1. Introduction

We consider an unknown function $\Phi : \mathcal{X} \times \Omega \to [0,1] \subset \mathbb{R}$, where $\mathcal{X} \subset [0,1]^D$ and Ω denotes the probability space representing some uncontrollable variables. For any fixed $x \in \mathcal{X}, Y_x = \Phi(x, \cdot)$ is a random variable of distribution $\mathbb{P}_x$ and we consider $g(x) = \psi(\mathbb{P}_x)$ with ψ a real-valued functional defined on probability measures. We assume that there exists at least one $x^* \in \mathcal{X}$ such that $g(x^*) = \sup_{x \in \mathcal{X}} g(x)$. Using a set of sequential observations $(\Phi(x_1, \omega_1), \cdots, \Phi(x_T, \omega_T))$, our goal is to minimize the simple regret $r_T = g(x^*) - g(x_T)$, with x_T the value returned after using a budget T.

Different families of algorithms have been developed to treat this problem. Some are for example of Bayesian flavor (see Shahriari et al., 2016, for instance), some are inspired by the bandit literature. Here we focus our interest on the bandit framework.

In the classical $\mathcal{X}$-armed bandit problem, a forecaster selects repeatedly a point x in the input space $\mathcal{X} \in [0,1]^D$ and receives a reward distributed according to an unknown distribution $\mathbb{P}_x$. Historically, the main goal was to minimize the *cumulative regret*, i.e. the sum of the difference between his collected rewards and the ones that would have been brought by optimal actions. In the last decade, other works focused on the simple regret. These

can be divided in two: algorithms that optimize an unknown function with the knownledge of the smoothness, for example StoOO (Munos et al., 2014), HOO (Bubeck et al., 2011) or Zooming (Kleinberg et al., 2008) and others focusing on the optimization of unknown functions without the knowledge of the smoothness, such as POO (Grill et al., 2015), StroquOOL (Bartlett et al., 2018), GPO (Xuedong et al., 2019), StoSOO (Valko et al., 2013) or Locatelli and Carpentier (2018).

Those algorithms focus on the optimization of the conditional *expectation* of $\mathbb{P}_x$. This choice is questionable in some situations. For example if the shape and variance of the reward distribution depend on the input, a forecaster may be interested in different aspects of the unknown distribution in order to modulate its risk exposure. In the literature, some measures of risk have been proposed to replace the expectation: for instance quantiles (also referred to as Value-at-Risk, see Artzner et al., 1999), the Conditional Value-at-Risk (CVaR also referred as Superquantile or Expected Shortfall, Rockafellar et al., 2000) or expectiles (Bellini and Di Bernardino, 2017). The purpose of this paper is to present a risk optimization framework of an unknown stochastic function with the knowledge of the smoothness using only pointwise sequential observations and a finite budget T.

$\mathcal{X}$-armed bandit algorithms rely on *optimistic strategies* that associate with each point of the space an upper confidence bound (UCB), that is, an *optimistic* prediction of the outcome. Adapting the classical setting to the optimization of risk measures implies being able to create high-probability confidence bounds for that particular measure. This problem has been tackled in the multi-armed bandit setting (*i.e.* when the input space is discrete and finite). For instance, Audibert et al. (2009); Sani et al. (2012) focused on the empirical variance, Galichet et al. (2013); Kolla et al. (2019); Hepworth (2017) on the CVaR while in David and Shimkin (2016); Szorenyi et al. (2015) the authors based their policies on the quantile. However, the literature is scarce in the continuous input space case.

In this paper we provide a new version of the Stochastic Optimistic Optimization (StoOO) algorithm (Munos et al., 2014), named StoROO (Stochastic Risk Optimistic Optimization), which is designed to handle any functional ψ. In a first part, we provide an analysis of the simple regret from a generic point of view. We then particularize our analysis in two important illustrative cases: conditional quantiles and CVaR. In the case of quantiles, assuming that the output distribution has a continuous, strictly increasing cumulative distribution function, we first propose an upper bound on the simple regret using Hoeffding's inequality, then, we derive tighter confidence intervals that take into account the order of the quantile respectively based on Bernstein's and Chernoff's inequalities. In the case of the CVaR, we first derive an upper bound on the regret using the deviation inequality of Brown (2007), then using the work of Thomas and Learned-Miller (2019) we derived tighter confidence bounds. Finally, we present numerical experiments that illustrate the ability of our method to optimize conditional quantiles and CVaR of a black-box function and the relevance of using tight deviation bounds. Due to space limitation, all proofs are deferred to Supplementary Material.

2. Problem setup

2.1. Hierarchical partitioning

The upper confidence bounds on which optimistic algorithms are based are surrogate functions $U : \mathcal{X} \to \mathbb{R}$ larger than the objective (in a sense detailed below) with high probability. At each round t, the point $X(t)$ having the highest UCB is sampled and a reward $Y_X(t)$ is collected.

In the classical multi-armed bandit problem, computing and sorting the UCB can be done without major issues. But dealing with continuous input spaces implies maximizing a UCB function over a continuous space, which can be both computational intensive and algorithmically challenging. For example, Piyavskii's algorithm (see Bouttier, 2017, and references therein) defines U using a global Lipschitz assumption on the targeted function. Because of the Lipschitz hypothesis, the UCB maximizer is at an intersection of hyperplanes, i.e. where the UCB is non-differentiable. Thus a gradient-based algorithm cannot be used, implying that finding the point with the highest UCB is a very hard problem to solve.

To overcome the computational difficulties, a popular alternative is to rely on hierarchical partitions (see Bubeck et al. (2011); Munos et al. (2014) for instance), $\mathcal{P} = \{\mathcal{P}_{h,j}\}_{h,j}$ of $\mathcal{X}$ such that

$$\mathcal{P}_{0,1} = \mathcal{X}, \quad \mathcal{P}_{h,j} = \bigcup_{i=0}^{K-1} \mathcal{P}_{h+1,Kj-i} \ ,$$

with K the number of sub-regions obtained after expanding a cell and $\mathcal{P}_{h,j}$ the j-th cell at depth h. In the following we assume that:

Assumption 1: There exists a decreasing sequence $\delta(h)$, such that for any $h \geq 0$ and for any cell $\mathcal{P}_{h,j}$, $\sup_{x \in \mathcal{P}_{h,j}} \|x - x_{h,j}\|_\infty \leq \delta(h)$, with $x_{h,j}$ the center of $\mathcal{P}_{h,j}$.

Assumption 2: There exists $\nu > 0$ such that every cell of depth h contains a ball of radius $\nu\delta(h)$.

Starting with $\mathcal{P}_{0,1}$ and following an optimistic strategy, at time t the algorithm has expanded some cells and the result is a tree $\mathcal{T}_t$ that is a subset of $\mathcal{P}$ and a partition of $\mathcal{X}$. In this setting U is taken as a piecewise constant function. Indeed for any $(\mathcal{P}_{h,j})_{h,j \in \mathcal{T}_t}$ we define $\bar{U}_{h,j}$ such that for all $x \in \mathcal{P}_{h,j}$, $U(x) = \bar{U}_{h,j}$.

In the literature of $\mathcal{X}$-armed bandits there are two ways to select a cell of $\mathcal{T}_t$ at each round. In Bubeck et al. (2011), the algorithm follows an *optimistic path* from the root to the leaves. In Munos et al. (2014), StoOO selects the cell having the highest UCB among all the cells of $\mathcal{T}_t$ that have not been expanded, *i.e.* the set $\mathcal{L}_t$ of leaves of $\mathcal{T}_t$. We consider here this second alternative. Hence, to find the maximizer of U at time t, we only need to evaluate and sort a finite number of values $(\bar{U}_{h,j})_{(h,j) \in \mathcal{L}_t}$.

2.2. Regularity assumptions, noise and bias

Even in the absence of noise, optimization from finite samples requires some regularity of the objective. Following Munos et al. (2014), we assume the following smoothness property:

$$\forall x \in \mathcal{X}, \quad g(x) \geq g(x^*) - \beta ||x - x^*||^\gamma \text{ with } \gamma, \ \beta \ > 0 \ . \tag{1}$$

Note that this condition is less restrictive than a global Hölder condition. In particular, the objective may be very irregular (even possibly discontinuous) except in the neighborhood of global maxima.

At first glance, in our stochastic setting, it may not be easy to asses that g satisfies (1). Sufficient conditions can be derived from the continuity of the conditional distribution $\mathbb{P}_x$ with respect to x. The relevant metric on the space of distributions actually depends on the chosen risk. For conditional quantiles, the natural assumption is that $x \mapsto F_x^{-1}(\tau)$ satisfies (1), and a sufficient condition is that $\|F_x^{-1} - F_y^{-1}\|_\infty \leq \beta\|x-y\|^\gamma$. In the case of the the conditional expectation and for the CVaR (or more generally for a large class of Optimized Certainty Equivalent Ben-Tal and Teboulle (2007)), the natural metric involved is the *Wasserstein distance* $\mathcal{W}_1$, as explained in Section 1 of the supplementary material.

To create confidence bounds for $(\mathcal{P}_{h,j})_{(h,j)\in\mathcal{L}_t}$, StoOO samples the leafs at their centers $(x_{h,j})_{(h,j)\in\mathcal{L}_t}$. Then using that all observed values are independent, *deviation inequalities* are used to create $(U_{h,j})_{(h,j)\in\mathcal{L}_t}$, a UCB for $(g(x_{h,j}))_{(h,j)\in\mathcal{L}_t}$. Finally to create $(\bar{U}_{h,j})_{(h,j)\in\mathcal{L}_t}$, a UCB over the cells, a *bias term* is added that takes into account how g can potentially increase from the center of the cell to its edges. Because the convergence of StoOO (and StoROO) only needs $\bar{U}_{h,j}$ to be a UCB of $\max_{x\in\mathcal{P}_{h,j}} g(x)$ for the cell containing x^* (see the proof of Proposition 2 (see also Munos et al. (2014)), it is enough to use the condition (1) to define a UCB as

$$\bar{U}_{h,j} = U_{h,j} + B_{h,j}, \text{ with } B_{h,j} = \hat{\beta}\delta(h)^{\hat{\gamma}} ,$$

and $\beta \leq \hat{\beta}$, $\gamma \geq \hat{\gamma}$. The algorithm also needs a quantity that bounds g from below in order to provide guaranties on the value of g over each cell. We thus construct a lower confidence bound, termed $L_{h,j}$, for $g(x_{h,j})$, and use it as a LCB for the maximum of g on $\mathcal{P}_{h,j}$. In particular, on the cell $\mathcal{P}_{h^*,j^*}$ containing the optimum x^*, it holds that

$$L_{h^*,j^*} \leq g(x^*) \leq U_{h^*,j^*} + \hat{\beta}\delta(h^*)^{\hat{\gamma}}$$

with high probability. To summarize, the estimation of $g(x^*)$ is altered by two sources of error: the local estimation error $E_{h^*,j^*} = U_{h^*,j^*} - L_{h^*,j^*}$ made at the center of the cell, and the bias term B_{h^*,j^*}. Balancing those two terms naturally provides a trade-off between exploration and exploitation.

3. Stochastic Risk Optimistic Optimization

3.1. The StoROO algorithm

StoROO starts by sampling one time each K sub-region of the root node. Then, at each time $1 \leq t \leq T$ the algorithm selects $\mathcal{P}_{h_t,j_t} \in (\mathcal{P}_{h,j})_{(h,j)\in\mathcal{L}_t}$ having the highest UCB. To reduce the estimation error, StoROO can either get more samples from $\mathcal{P}_{h_t,j_t}$ (to reduce the variance), or split the cell in order to reduce its diameter (to reduce the bias). The good balance between these two options is found by dividing a cell as soon as the local estimation error is smaller than the bias, that is when

$$U_{h_t,j_t} - L_{h_t,j_t} \leq \hat{\beta}\delta(h_t)^{\hat{\gamma}} . \tag{2}$$

If Condition (2) is satisfied, StoROO expands $\mathcal{P}_{h_t,j_t}$ and requires a new sample at the center of each sub-region. If Condition (2) is not satisfied, then StoROO requires a new sample at the center x_{h_t,j_t} which is used to update U_{h_t,j_t} and L_{h_t,j_t}.

When the budget is exhausted, several choices are possible for the return value: they have the same theoretical guarantees. Following Munos et al. (2014), one can return the deepest node among those that have been expanded. Here we propose a different, more conservative choice. Denoting by $\mathscr{L}_T$ the set of nodes having the highest LCB among those that have been expanded after a budget T, StoROO returns the node with the highest value $\hat{g}$ (an estimator of g) among the deepest nodes of $\mathscr{L}_T$. The pseudo-code of the full algorithm is given in Algorithm 1. It requires the parameters $\hat{\beta}$ and $\hat{\gamma}$ that satisfy Condition (1), but of course the inequality do not have to be tight.

Algorithm 1 StoROO

Input: error probability $\eta > 0$; number of children K; time horizon T; $\hat{\beta} > 0$; $\hat{\gamma} > 0$;
Define: UCB and LCB
Initialization $n = 1$; $t = 1$;
Expand into K sub-regions the root node $(0, 0)$ and sample one time each child
while $n \leq T$ **do**
  foreach $(h, j) \in \mathcal{L}_t$ **do**
    compute $\bar{U}_{h,j}(t)$
  end
  Select $(\tilde{h}, \tilde{j}) = \arg\max_{(h,j)\in\mathcal{L}_t} \bar{U}_{h,j}(t)$
  Compute the LCB $L_{\tilde{h},\tilde{j}}(t)$
  if $U_{\tilde{h},\tilde{j}}(t) - L_{\tilde{h},\tilde{j}}(t) \leq \hat{\beta}\delta(h)^{\hat{\gamma}}$ **then**
    expand the node, remove $(\tilde{h}, \tilde{j})$ from $\mathcal{L}_t$, add to $\mathcal{L}_t$ the K sub-cells of $\mathcal{P}_{\tilde{h},\tilde{j}}$ and sample each new node once,
    $n = n + K$, $t = t + 1$
  else
    Sample the state $x_t = x_{\tilde{h},\tilde{j}}$ and collect the observation $Y_{x_{h_t,j_t}}$, $n = n + K$, $t = t + 1$
  end
end
Return the node according to the returning rule.

3.2. Analysis of the algorithm

In this section we provide a theoretical analysis of StoROO. It is inspired by Munos et al. (2014), but differs most notably by the fact that the analysis is suited for any g and not only for the conditional expectation. The analysis relies on the possibility to construct, for any $\eta > 0$, upper- and lower-confidence bounds $U^{\eta}_{h,j}(t)$ and $L^{\eta}_{h,j}(t)$ such that the event

$$\mathcal{A}_\eta = \bigcap_{T\geq t\geq 1} \bigcap_{\mathcal{P}_{h,j}\in\mathcal{T}_t} \left\{ U^{\eta}_{h,j}(t) \geq g(x_{h,j}),\ L^{\eta}_{h,j}(t) \leq g(x_{h,j}) \right\}$$

has probability $\mathbb{P}(\mathcal{A}_\eta)$ at least $1 - \eta$. We defer to Section 4 their specific expression for the cases of the quantile and CVaR. Especially Section 4 shows that in our setting the size of the confidence interval associated to each node is not always explicit, by opposition of the classical case. We thus need to introduce the following definition to quantify how many times a node needs to be sampled before satisfying the expansion condition (Eq. 2).

Definition 1 *Let* $m_{\eta,h}(\theta,\kappa,\alpha) = \log(\theta T^2/\eta)\left(\frac{\kappa}{\hat{\beta}\delta(h)^{\hat{\gamma}}}\right)^{\alpha}$ *and* $N_{h,j}(t) = \sum_{s=1}^{t} \mathbb{1}_{X(s)\in\mathcal{P}_{h,j}}$, *a* vector of safe constants $v = (\theta,\kappa,\alpha)$ *is composed of constants* $\theta > 0$, $\kappa > 0$, *and* $\alpha > 0$ *such that the event*

$$\mathcal{B}_\eta = \bigcap_{T\geq t\geq 1} \bigcap_{N_{h,j}\geq m_{\eta,h}(\theta,\kappa,\alpha)} \bigcap_{\mathcal{P}_{h,j}\in\mathcal{T}_t} \left\{ U^{\eta}_{h,j}(t) - L^{\eta}_{h,j}(t) \leq \hat{\beta}\delta(h)^{\hat{\gamma}} \right\}$$

has probability at least $1-\eta$.

For example, in the case of the conditional expectation a direct consequence of Hoeffding's inequality provides $\theta = 2$, $\alpha = 2$ and $\kappa = \sqrt{1/2}$ (see Munos et al. (2014)).

To ensure the convergence of StoROO, we first prove (Proposition 2) that any point at the center of an expanded cell of depth h belongs to

$$J_h = \{\ x_{h,j} \text{ such that } g(x_{h,j}) + 2\hat{\beta}\delta(h)^{\hat{\gamma}} \geq g^*\}\ . \tag{3}$$

Next, Proposition 3 shows that using a budget T, the tree $\mathcal{T}_T$ reaches at least a depth $H^*_\eta(T)$. This implies the point returned by the algorithm belongs to $J_{H^*_\eta(T)}$ (Proposition 4). Finally, using an assumption on the size of J_h that can be formalized by the so-call *near-optimality dimension*, we provide an upper bound on the regret (Theorem 7).

Proposition 2 *Conditionally on* $\mathcal{A}_\eta$, *StoROO only expands cells* $\mathcal{P}_{h,j}$ *such that* $x_{h,j} \in J_h$.

Given the safe constants v and the total budget T, the deeper the algorithm builds the tree, the better are the guarantees on the final point returned. So the goal of the following proposition is to provide a lower bound on the depth of $\mathcal{T}_T$.

Proposition 3 *Define* $n_{\eta,h} = m_{\eta,h}(v)$ *and define* H_η *the largest* $h \in \mathbb{N}$ *such that*

$$S_h = K\sum_{h'\leq h} n_{\eta,h'+1}|J_{h'}| \leq T, \quad \textit{with} \quad |J_{h'}| \textit{the cardinal of } J_{h'}\ .$$

The deepest node H^*_η *expanded by StoROO is such that* $H^*_\eta \geq H_\eta$.

Intuitively, S_h is the budget needed to expand all the nodes in J_h for all $h' \leq h$. It may be that some of this nodes will not be visited, but in the worst case they are and they need to be considered in order to obtain a valid bound. Putting Propositions 2 and 3 together, yields a first upper bound on the simple regret:

Proposition 4 *Running StoROO with budget* T, *with probability* $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta)$ *the regret is bounded as*

$$r_T \leq 2\hat{\beta}\delta\big(H^*_\eta(T)\big)^{\hat{\gamma}}\ .$$

A more explicit bound for the regret can be obtained by quantifying the volume of $\mathcal{X}_\epsilon = \{x \in \mathcal{X},\ g(x) \leq g^* - \epsilon\}$ for small values of ϵ. Introducing the Holderian semi-metric $\ell_{\beta,\gamma}(x,x') = \beta\,\|x-x'\|^{\gamma}$, that is associated with its regularity constants β and γ, the *near-optimality* dimension of the function is defined as follows, (see Munos et al. (2014); Bubeck et al. (2011) for more details).

Definition 5 *The ν-near optimality dimension is the smallest $d \geq 0$ such that for all $\epsilon \geq 0$, there exists $C \geq 0$ such that the maximal number of disjoint $\ell_{\hat{\beta},\hat{\gamma}}$-balls of radius $\nu\epsilon$ with center in $\mathcal{X}_\epsilon$ is less than $C\epsilon^{-d}$.*

In order to evaluate H_η^*, we need to bound $|J_h|$ for all $h \geq 0$. The following proposition makes the link between the near optimality dimension and $|J_h|$.

Proposition 6 *Let d be the $\frac{\nu^{\hat{\gamma}}}{2}$-near-optimality dimension, and C the corresponding constant. Then*

$$|J_h| \leq \frac{C}{\left(2\hat{\beta}\delta(h)^{\hat{\gamma}}\right)^d} .$$

Finally, combining Propositions 4 and 6 with an hypothesis on the decreasing sequence $\delta(h)$, it is possible to provide the speed of convergence of r_T.

Theorem 7 *Assume that $\delta(h) = c\rho^h$ for some $c \geq 0$ and $\rho < 1$, and assume that $v = (\theta, \kappa, \alpha)$. Thus with probability $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta)$, the regret of StoOO is bounded as*

$$r_T \leq c_1 \left[\frac{\log(\theta T^2/\eta)}{T}\right]^{\frac{1}{d+\alpha}} \quad \text{with} \quad c_1 = 2\hat{\beta}\left[\frac{KC\kappa^\alpha [2\hat{\beta}]^{-d}}{(1-\rho^{d\hat{\gamma}+\hat{\gamma}\alpha})}\right]^{\frac{1}{d+\alpha}},$$

where d is the near optimality dimension and C the corresponding near optimality constant.

The proof is deferred to Supplementary Material. If g is the conditional expectation, a vector of *safe constants* is $(\theta = 2, \alpha = 2, \kappa = \sqrt{1/2})$ (based on Hoeffding's inequality). Thus if we plug it into the quantity defined in Theorem 7 we obtain

$$r_T \leq c_1 \left[\frac{\log(2T^2/\eta)}{T}\right]^{\frac{1}{d+2}} \quad \text{with} \quad c_1 = 2\hat{\beta}\left[\frac{KC[2\hat{\beta}]^{-d}}{2(1-\rho^{d\hat{\gamma}+\hat{\gamma}\alpha})}\right]^{\frac{1}{d+2}},$$

that is equivalent to what it is obtained in Munos et al. (2014).
Remark: In the particular case where each cell is a hypercube and the sub-regions are created by the division of the parent-cell into $K = 2^D$ sub-regions of equal size, then $K = 2^D$, c is equal to $\sqrt{D}$ and ρ is equal to $\frac{1}{2}$.

4. Optimizing quantiles

In this section, we focus on the optimization of *quantiles*, which are well-established tools in (risk-averse) decision theory (see Rostek, 2010, for instance). In particular, they benefit from interesting robustness properties, with respect to outliers or heavy tails. Let

$$g(x) = q_x(\tau) = \inf\left\{q \in \mathbb{R} : F_x(q) \geq \tau\right\},$$

be the τ-quantile of Y_x, where F_x is the cumulative distribution function (CDF) of $\mathbb{P}_x$. Here we detail how to construct the UCB and LCB for quantiles. First, we provide bounds based on Hoeffding's inequality and we use them to adapt the regret bounds of Theorem 7. Then we provide two more refined bounds that take into account the order τ of the quantile based respectively on the Bernstein's inequality and on the Kullback-Leibler divergence.

Let us first introduce some notations. For all $1 \leq t \leq T$, $1 \leq h \leq t$, $1 \leq j \leq K^h$ and $q \in \mathbb{R}$ we denote

$$\hat{F}^t_{h,j}(q) = \frac{\sum_{s=1}^t \mathbb{1}_{Y(t)\leq q} \mathbb{1}_{X(t)\in\mathcal{P}_{h,j}}}{N_{h,j}(t)},$$

the empirical CDF of the reward inside the cell $\mathcal{P}_{h,j}$, where $N_{h,j}(t)$ is the (random) number of times the cell was sampled up to time t (see Definition 1). The *generalized inverse* $\hat{F}^{t\ -}_{h,j}$ of the piecewise constant function $\hat{F}^t_{h,j}$ is defined as $\hat{q}_{h,j}(\tau) = \inf\left\{q \in \mathbb{R} :\ \hat{F}^t_{h,j}(q) \geq \tau\right\}$, that is the $\lceil N_{h,j}(t) \times \tau \rceil$ order statistic of the sample that has been collected from the node $x_{h,j}$ until time t.

To define confidence bounds on the conditional quantile we proceed in two steps. First we propose confidence bounds on $\hat{F}_{h,j}(q_\tau)$. To do so, we simply use deviation bounds for Bernoulli distributions, since for all $x \in \mathcal{X}$, for all $1 \leq n \leq T$, the random variables $\left(\mathbb{1}_{Y_x(\xi_s)\leq q_x(\tau)}\right)_{s=1,\cdots,n}$ are independent and identically distributed with a Bernoulli law of parameter τ, if ξ_s denotes the time when the node x has been sampled for the s-th time. Then we use the properties

$$\forall\ \epsilon > 0\ \text{ such that }\ \tau + \epsilon < 1, \quad \hat{F}^t_{h,j}(q_{h,j}(\tau)) \geq \tau + \epsilon \quad \Leftrightarrow \quad q_{h,j}(\tau) \geq \hat{F}^{t\ -}_{h,j}(\tau + \epsilon)\ , \quad (4)$$

$$\forall\ \epsilon > 0\ \text{ such that }\ \tau + \epsilon > 0, \quad \hat{F}^t_{h,j}(q_{h,j}(\tau)) < \tau - \epsilon \quad \Leftrightarrow \quad q_{h,j}(\tau) \leq \hat{F}^{t\ -}_{h,j}(\tau - \epsilon)\ , \quad (5)$$

to create confidence bounds on $q_{h,j}(\tau)$ using bounds on $\hat{F}^t_{h,j}(q_\tau)$. Note that here we just assume that the output distribution has a continuous, striclty increasing cumulative distribution function. It is not necessary to assume something else, such as bounded support or bounded moments because here we refer to Bernouilli distributions. The first equivalence in illustrated on Figure 1 of the supplementary material.

4.1. Hoeffding's bound and regret analysis

Let $\epsilon^{\eta,T}_{N_{h,j}(t)} = \sqrt{\dfrac{\log(2T^2/\eta)}{2N_{h,j}(t)}}$, and let

$$U^\eta_{h,j}(t) = \begin{cases} \min\left\{q,\ \hat{F}^t_{h,j}(q) \geq \tau + \epsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau + \epsilon^{\eta,T}_{N_{h,j}(t)} < 1 \\ +\infty & \text{otherwise,} \end{cases} \quad (6)$$

$$L^\eta_{h,j}(t) = \begin{cases} \max\left\{q,\ \hat{F}^t_{h,j}(q) \leq \tau - \epsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau - \epsilon^{\eta,T}_{N_{h,j}(t)} > 0 \\ -\infty & \text{otherwise.} \end{cases} \quad (7)$$

The next proposition motivates the choice of the above quantities as a UCB and a LCB for the quantile of order τ at the points $(x_{h,j})_{(h,j)\in\mathcal{T}_t}$.

Proposition 8 *Assume that for all $x \in \mathcal{X}$, $\mathbb{P}_x$ has a continuous, striclty increasing cumulative distribution function then for any $\eta > 0$, for all $h \geq 0$, for all $0 \leq j \leq K^h$ and for all $1 \leq t \leq T$, if $L^\eta_{h,j}(t)$ and $U^\eta_{h,j}(t)$ are defined according to (7) and (6), respectively, then the event $\mathcal{A}_\eta$ has probability at least $1 - \eta$.*

The proof is deferred to Supplementary Material. Now, analyzing the regret requires a high probability bound on the number of time a node is sampled before being expanded:

Proposition 9 *Under the conditions required by Proposition 8, define f_x as the density of $\mathbb{P}_x$ and define $\bar{f}(x) = \min_{\tau' \in [\tau - 2\epsilon^{\eta,T}_{M_\tau}, \tau + 2\epsilon^{\eta,T}_{M_\tau}]} f_x \circ F_x^{-1}(\tau')$ with $M_\tau = 2m_\tau^{-2}\log(2T^2/\eta)$ and $m_\tau = \min(\tau, 1-\tau)$. If $U^\eta_{h,j}(t)$ and $L^\eta_{h,j}(t)$ are defined according to (6) and (7), respectively, then for any $\eta > 0$, $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta) \geq 1 - \eta$ and a vector of safe constants is given as*

$$v = \left(2, \frac{\sqrt{8m_\tau^2 + 4\big(\hat{\beta}\,\mathrm{diam}(\mathcal{X})^{\hat{\gamma}} \min_{x\in\mathcal{X}} \bar{f}(x)\big)^2}}{m_\tau \min_{x\in\mathcal{X}} \bar{f}(x)}, 2\right).$$

The proof is deferred to Supplementary Material. According to the previous proposition, if we have sampled a node at depth h more than

$$n_{\eta,h} = \log(2T^2/\eta)\left(\frac{8m_\tau^2 + 4\big(\hat{\beta}\,\mathrm{diam}(\mathcal{X})^{\hat{\gamma}} \min_{x\in\mathcal{X}} \bar{f}(x)\big)^2}{\big(\min_{x\in\mathcal{X}} \bar{f}(x) m_\tau \hat{\beta}\delta(h)^{\hat{\gamma}}\big)^2}\right) \tag{8}$$

times, then with probability $1-\eta$, Condition (2) is satisfied and thus the node is expanded.

Equality (8) reflects two dependencies. The smaller the minimum of the density over a neighborhood of the quantile and the closer τ from 0 or 1, the larger the upper bound on the number of samples needed before being expanded. Indeed a small density value in a neighborhood of the targeted quantile will produce samples with few observations close to the quantile, hence the estimation error will be large. In addition from Proposition (8), to obtain non trivial UCB and LCB, the value $N_{h,j}$ has to be large enough to ensure $\tau \pm \epsilon^{\eta,T}_{N_{h,j}} \in [0,1]$ and this value increases as τ comes close from 0 or 1. Thus a more precise way to understand the behaviour of StoROO is that the number of time a node needs to be sampled before expansion depends on the pdf value in a neighborhood (of decreasing size with $N_{h,j}$) of the targeted quantile.

To obtain an upper bound on the simple regret, we now just need to combine Theorem 7 with Proposition 9 so as to obtain the following theorem.

Theorem 10 *Under the conditions required by Proposition 8 and 9, if $\delta(h) = c\rho^h$ for some $c \geq 0$ and $\rho < 1$, then with probability $1-\eta$, the regret of StoROO for maximizing the quantile is bounded as*

$$r_T \leq c_2 \left[\frac{\log(2T^2/\eta)}{T}\right]^{\frac{1}{d+2}} \text{ with } c_2^{d+2} = KC\hat{\beta}^2 \frac{16m_\tau^2 + 8\big(\hat{\beta}\,\mathrm{diam}(\mathcal{X})^{\hat{\gamma}} \min_{x\in\mathcal{X}} \bar{f}(x)\big)^2}{\big(m_\tau \min_{x\in\mathcal{X}} \bar{f}(x)\big)^2 (1 - \rho^{d\hat{\gamma}+\hat{\gamma}\alpha})},$$

with d the near-optimality dimension and C the near-optimality corresponding constant.

Note that the speed of convergence is the same as the one obtained in the conditional expectation optimization setting; only the constant varies.

4.2. Tighter bounds

Using Hoeffding's inequality is convenient because it leads to explicit lower and upper confidence bounds, which simplifies the deriviation of bounds on the regret. However, it implicitly upper-bounds the variance of all $[0,1]$-valued random variables by $1/4$, which is overly pessimistic when the inequality is applied to variables whose expectations are far from $1/2$.

This is in particular the case for quantile estimation, when the quantile is of order close to 0 or 1. To take into account the order of the quantile, following David and Shimkin (2016), a first possibility is to derive confidence intervals from Bernstein's inequality as presented in Proposition 1 of the supplementary material.

Although Bernstein's inequality takes into account the order of the quantile, it is possible to do something better. In order to create tighter confidence bounds, we thus go back to Chernoff's inequality and derive less explicit, but more accurate upper- and lower- confidence bounds on the τ-quantiles. We follow here Garivier and Cappé (2011), but a close inspection at the proofs shows however a difference in the order of the marginals of the KL functions. Recall that the binary relative entropy is defined for $(p,q) \in [0,1]^2$ as:

$$\mathrm{kl}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \, ,$$

with by convention, $0 \log 0 = 0$, $\log 0/0 = 0$ and $x \log x/0 = +\infty$ for $x > 0$.

Proposition 11 *For any $\eta > 0$, for all $1 \le t \le T$, $1 \le h \le t$ and $1 \le j \le K^h$, define*

$$U^\eta_{h,j}(t) = \min \left\{ q, \ \hat{F}^n_{h,j}(q) \ge \tau \text{ and } \mathrm{kl}(\hat{F}^t_{h,j}(q), \tau) \ge \frac{\log(2T^2/\eta)}{N_{h,j}(t)} \right\} \text{ if } \mathrm{kl}(1,\tau) > \frac{\log(2T^2/\eta)}{N_{h,j}(t)}$$

and $+\infty$ otherwise. Define

$$L^\eta_{h,j}(t) = \max \left\{ q, \ \hat{F}^t_{h,j}(q) \le \tau \text{ and } \mathrm{kl}(\hat{F}^t_{h,j}(q), \tau) \ge \frac{\log(2T^2/\eta)}{N_{h,j}(t)} \right\} \text{ if } \mathrm{kl}(0,\tau) > \frac{\log(2T^2/\eta)}{N_{h,j}(t)}$$

and $-\infty$ otherwise. Then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.

The proof is deferred to Supplementary Material. Contrary to Bernstein's inequality, Chernoff's bound is always tighter than Hoeffding's inequality, which follows from Pinsker's inequality (see e.g. Garivier et al., 2018). It follows in particular that the regret of StoROO using confidence bounds derived from Chernoff's inequality has, at least, the guarantees presented in Theorem 10.

The online setting we consider in this article induces that, after t steps, the set of nodes and the number of observations in each node are random. To cope with this, we thus need deviation bounds for random size samples. The most simple way to obtain such inequalities is to use a union bound on the possible number of observations in each node, as presented above. Tighter results can be obtained from a more thorough analysis (sometimes called *peeling trick*): this is what is presented below.

Proposition 12 *For any $\eta \in (0,1)$ let $\delta_\eta(T) = \inf \left\{ \delta > 0 : \ Te\lceil \delta \log(T) \rceil \exp(-\delta) \le \eta/2 \right\}$, and define*

$$U^\eta_{h,j}(t) = \min \left\{ q, \ \hat{F}^n_{h,j}(q) \ge \tau \text{ and } N_{h,j}(t) \, \mathrm{kl}(\hat{F}^t_{h,j}(q), \tau) \ge \delta_\eta(T) \right\} \text{ if } \mathrm{kl}(1,\tau) > \frac{\delta_\eta(T)}{N_{h,j}(t)}$$

and $+\infty$ otherwise. Define

$$L^\eta_{h,j}(t) = \max \left\{ q, \ \hat{F}^t_{h,j}(q) \le \tau \text{ and } N_{h,j}(t) \, \mathrm{kl}(\hat{F}^n_{h,j}(q), \tau) \ge \delta_\eta(T) \right\} \text{ if } \mathrm{kl}(0,\tau) > \frac{\delta_\eta(T)}{N_{h,j}(t)}$$

and $-\infty$ otherwise. Then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.

The proof is deferred to supplementary material. Note that for every $0 < \delta \le \log(2/\eta)$, $\lceil \delta \log(T) \rceil \ge 1$ and thus $Te\lceil \delta \log(T) \rceil \exp(-\delta) > \eta/2$; hence, $\delta_\eta(T) > \log(2/\eta)$.

5. Optimizing CVaR

We now detail how StoROO can be applied to the optimization of another important notion of risk: the CVaR. CVaR has raised a great interest in recent years, notably because it is a *coherent* risk indicator (see Ben-Tal and Teboulle (2007) for instance). For $\tau \in [0,1)$ the condition value at risk at level τ of a continuous random variable Y is defined as

$$\mathrm{CVaR}_\tau(Y) = \inf_{z\in\mathbb{R}} \left\{ z + \frac{1}{(1-\tau)}\mathbb{E}[(Y-z)^+] \right\} = \mathbb{E}\Big(Y | Y \geq q(\tau)\Big) \,,$$

with $(z)^+ = \max(0,z)$. Following Brown (2007), it can be estimated by

$$\widehat{\mathrm{CVaR}}_\tau^{\,n} = \inf_{z\in\mathbb{R}} \left\{ z + \frac{1}{(1-\tau)n}\sum_{i=1}^n (Y_i - z)^+ \right\} = Y_{(\lfloor n\tau \rfloor)} + \frac{1}{(1-\tau)n}\sum_{i=1}^n (Y_i - Y_{(\lfloor n\tau \rfloor)})^+ \,.$$

Since Y often stands for a loss, the CVaR is usually to be minimized. In order to stay consistent with the rest of the paper, we choose in the following to maximize $g = -\mathrm{CVaR}_\tau$.

Assuming the random variables are bounded in an interval $[a,b]$, the next proposition adapts the deviation inequalities presented in Brown (2007) to our sequential setting.

Proposition 13 *For any $\eta > 0$, for all $h \geq 0$, for all $0 \leq j \leq K^h$ and for all $1 \leq t \leq T$, define*

$$U^\eta_{h,j}(t) = -\widehat{\mathrm{CVaR}}_\tau^{\,n} + \frac{b-a}{1-\tau}\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}(t)}}, \quad L^\eta_{h,j}(t) = -\widehat{\mathrm{CVaR}}_\tau^{\,n} - (b-a)\sqrt{\frac{5\log(6T^2/\eta)}{(1-\tau)N_{h,j}(t)}}.$$

If the random variables Y_x are bounded in $[a,b]$ for all $x \in \mathcal{X}$ and have continuous distribution functions, then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.

Note that *deviation inequalities* can be established for CVaR in sub-Gaussian or light-tailed cases (see Kolla et al. (2019) for instance) but an assumption has to be made on the value of the pdf in a neighborhood of the τ-quantile.

From Proposition (13), one can see that whenever a node has been played more than $m_{\eta,h} = \log(6T^2/\eta)(b-a)^2\left(\frac{1+\sqrt{10(1-\tau)}}{\sqrt{2}(1-\tau)\hat{\beta}\delta(h)^{\hat{\gamma}}}\right)^2$ times, it has been expanded. Thus a possible associated vector of *safe constants* is $v = \left(6, (b-a)\left(\frac{1+\sqrt{10(1-\tau)}}{\sqrt{2}(1-\tau)\hat{\beta}\delta^{\hat{\gamma}}}\right), 2\right)$. Combining v with Theorem 7 provides the following upper bound on the regret.

Theorem 14 *Under the conditions required by Proposition 13, if $\delta(h) = c\rho^h$ for some $c \geq 0$ and $\rho < 1$, then with probability $1-\eta$, the regret of StoROO for minimizing CVaR_τ is bounded as*

$$r_T \leq c_3 \left[\frac{\log(6T^2/\eta)}{T}\right]^{\frac{1}{d+2}} \quad \textit{with} \quad c_3 = 2\hat{\beta}\left[\frac{\left(1+\sqrt{10(1-\tau)}\right)^2 KC(b-a)^2[2\hat{\beta}]^{-d}}{2(1-\tau)^2(1-\rho^{d\hat{\gamma}+\hat{\gamma}\alpha})}\right]^{\frac{1}{d+2}},$$

with d the near-optimality dimension and C the near-optimality corresponding constant.

The inequalities obtained in Proposition 13 are convenient because they lead to explicit lower and upper confidence bounds, which simplifies the derivation of bounds on the regret. However, as they are based on Hoeffding's inequality, they can be over-conservative. To obtain better bounds, Thomas and Learned-Miller (2019) propose data-dependent inequalities derived from the *Dvoretzky-Kiefer-Wolfowitz* inequality. Due to space limitation, the values of the UCB and LCB derived from Thomas and Learned-Miller (2019) are deferred to the Proposition 2 of the supplementary material. Although we do not propose an analysis of the regret based on this bounds, it is immediate to state that the upper bound on the regret is always smaller than the bound obtained in Theorem 14 because these inequalities are strictly tighter than Brown's inequalities. In the following section, we numerically highlight the relevance of using these tight bounds.

6. Experiments

We empirically highlight the capacity of StoROO to optimize the conditional quantile and CVaR of a black-box function. Four versions of StoROO are compared for both cases.

For the conditional quantile we compare StoROO using confidence bounds repectively derived from Hoeffding's, Bernstein's, Chernoff's inequalities (resp. denoted StoROO_{H}, StoROO_{B} and $\text{StoROO}_{\text{kl}}$) and Chernoff's inequality and the *peeling trick* ($\text{StoROO}_{\text{kl-p}}$).

For the optimization of the conditional CVaR, we compare the use of confidence bounds derived from Brown's inequality and from Thomas and Learned-Miller (2019). To use these inequalities we have to provide $(a, b) \in \mathbb{R}^2$ that bound the output. Hence, we compare two cases: one where we provide conservative bounds for (a, b) (here $(a, b) = (0, 1)$), and one where we provide their actual values ($a_x = \min \text{supp}(Y_x)$ and $b_x = \max \text{supp}(Y_x)$, *i.e.* the minimum and the maximum of the support of the conditional distribution). We denote the four variants $\text{StoROO}_{\text{Br}}$ (from Brown's inequality), StoROO_{T} (from Thomas and Learned-Miller (2019)), and $\text{StoROO}_{\text{Br-o}}$ and $\text{StoROO}_{\text{T-o}}$ for their variants with oracle bounds.

As a test-case, we chose two functions with heteroscedastic noise and local extrema. The first is $\Phi_1(x, \cdot) = 0.18(\sin(3x)\sin(13x) + 1.3) + 0.062\zeta(\cdot)\big(\cos(8x - 2) + 1.2\big)$, where ζ is a log-normal random variable of parameters 0 and 1 truncated at its 0.95-quantile (the truncated mass is uniformly reallozcated between $q(0.91)$ and $q(0.95)$). Note that to initialise StoROO not too close from a global optimum, we optimize the quantiles of Φ_1 on $[-0.1, 0.9]$ and the CVaR on $[0, 1]$. Figure 1 (left) shows the shape of the 0.1 and 0.9 -quantiles and -CVaR of Φ_1, while Figure 1 (right) shows samples of the 0.1-quantile. The second test-case is $\Phi_2(x, \cdot) = \text{Cr}(x) + \zeta(\cdot)|\,\text{Cr}(x) + 1.5\sqrt{x_1^2 + x_2^2}|$, on $[-0.5, 1]^2$ with

$$\text{Cr}(x) = 0.1\Big(\Big|\sin(x_1)\sin(x_2)\exp\Big(\Big|3 - (\sqrt{x_1^2 + x_2^2}/\pi)\Big|\Big)\Big| + 1\Big)^{1.4}$$

and ζ a random variable that follows a Cauchy distribution of parameters $(0, 0.75)$. Note that for all $x \in \mathcal{X}$, $\Phi_2(x, \cdot)$ is unbounded and it has unbounded moments. Thus we can only apply quantile optimization on Φ_2 based on the strategies developed in the past sections. Figure 2 (left) shows the shape of the 0.1-quantile of Φ_2. The performance of each version of StoROO is evaluated for different values of τ and quantified according to the simple regret. In our experiments we fix the values $\hat{\beta} = 12$ and $\hat{\gamma} = 1.4$ (resp. $\hat{\beta} = 2$, $\hat{\gamma} = 0.5$ and $\hat{\beta} = 2$, $\hat{\gamma} = 0.7$) for the optimization of the quantiles (resp. the CVaR of order 0.1 and 0.9) of Φ_1

and $\hat{\beta} = 13$ and $\hat{\gamma} = 1$ for the optimization of the 0.1-quantile of Φ_2. Note that these values underestimate the regularity conditions at optimum so that satisfying the condition (1). In addition we fix $K = 3^D$ and we choose to expand the nodes into sub-region of equal sizes.

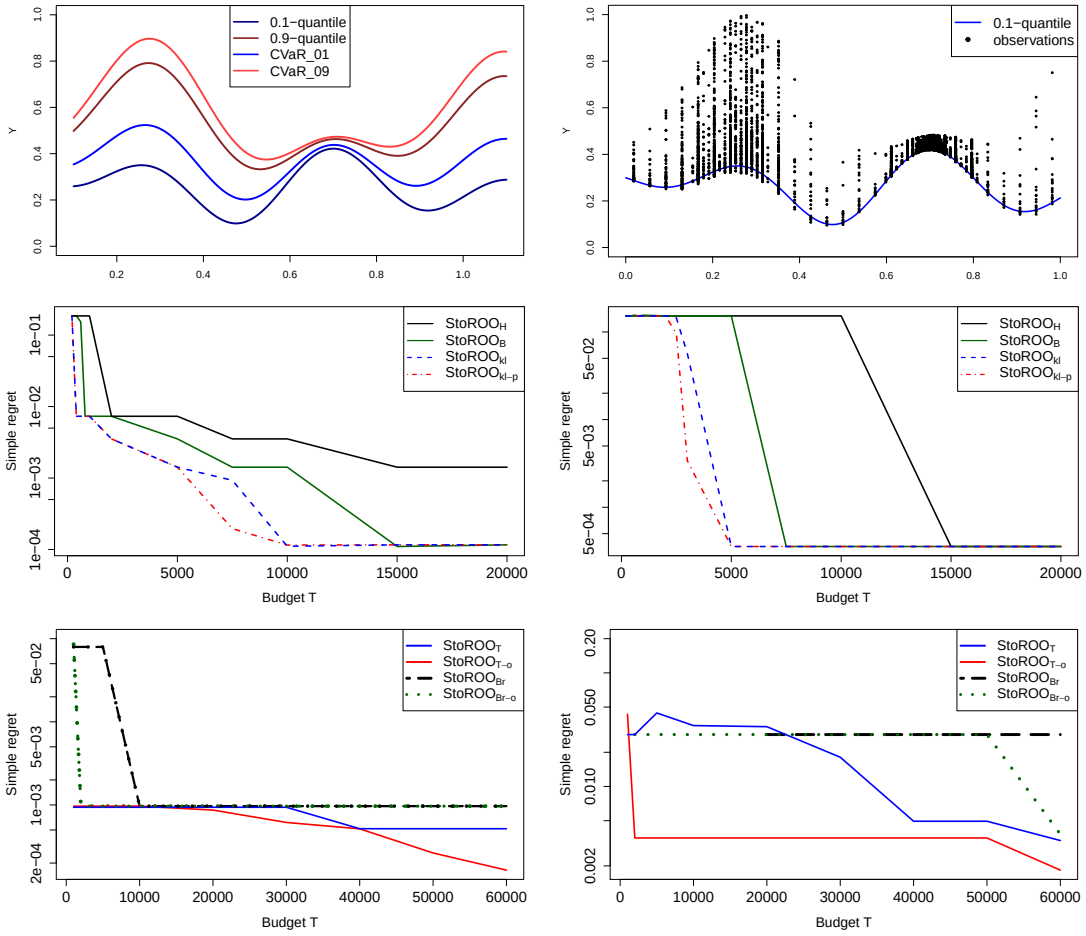

Figure 1: Results for the Φ_1 test function. Top left: conditional quantiles and CVaR of Φ_1. Top right: one run of StoROO$_{\text{kl}}$ for the 0.1-quantile with $T = 5,000$, $\hat{\beta} = 12$ and $\hat{\gamma} = 1.4$. Middle: evolution of the simple regret for the optimization of the quantile of order 0.1 (left) and 0.9 (right). Bottom: evolution of the simple regret for the optimization of the CVaR of order 0.1 (left) and 0.9 (right).

Figure 1 and 2 report the average of the simple regret over 100 runs. For both values of τ all the variants of StoROO have a regret that decreases with the budget. However from our experiments a ranking can be created. For the optimization of the quantile let us firt remark that as bounds are known for Φ_1, for this test case we modified Proposition (8-11-12) by replacing $(-\infty, +\infty)$ by $(0, 1)$. The less efficient method is StoROO$_{\text{H}}$. For $\tau = 0.9$ its simple regret decreases slower than the three others methods and for $\tau = 0.1$ StoROO$_{\text{H}}$ does not reach the performance of the others variants. To reach a fixed accuracy,

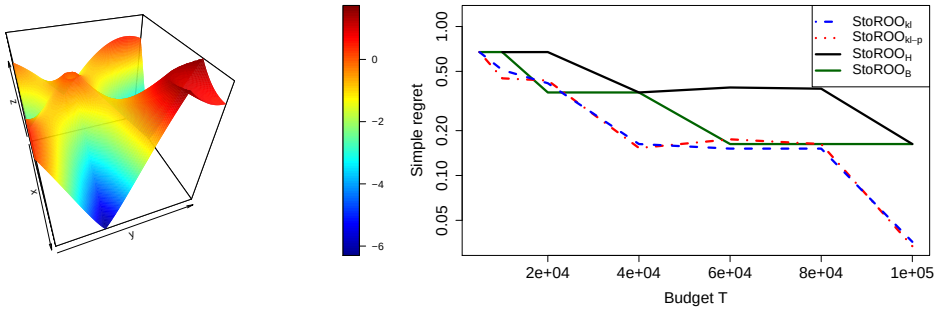

Figure 2: Results for the Φ_2 test function. Left: Conditional quantile of order 0.1 of Φ_2. Left: Simple regret for the optimization of the conditional quantile presented to the left.

StoROO$_{\text{H}}$ sometimes needs a much larger budget than others variants. For example, on Φ_1, taking $\tau = 0.9$, StoROO$_{\text{H}}$ needs a budget of $15,000$ to reach a simple regret of order 10^{-4}, while StoROO$_{\text{kl}}$ and StoROO$_{\text{kl-p}}$ need a budget equal to $5,000$. Second-to-last is StoROO$_{\text{B}}$. Using the maximal budget, on both experiments on Φ_1, this variant reaches the same accuracy as StoROO$_{\text{kl}}$ and StoROO$_{\text{kl-p}}$ but its simple regret decreases slower. For some levels of performance StoROO$_{\text{B}}$ needs a much larger budget than StoROO$_{\text{kl}}$. For example, taking $\tau = 0.1$, to reach the value $r_T = 10^{-4}$ StoROO$_{\text{B}}$ needs a budget of $T = 15,000$ while $T = 10,000$ is enough for StoROO$_{\text{kl}}$. Finally, the most efficient methods are clearly StoROO$_{\text{kl}}$ and StoROO$_{\text{kl-p}}$. The use of a peeling argument (instead of a plain union bound) in StoROO$_{\text{kl-p}}$ provides some additional gain over StoROO$_{\text{kl}}$ on Φ_1 but the effect is negligible on Φ_2.

For the optimization of the CVaR, the variant based on tighter bounds is almost always better than the other and it is independent of the use of oracle bounds. The use of oracle bounds always improves the performance of StoROO and this effect is stronger if the confidence intervals are created with the inequalities of Thomas and Learned-Miller (2019). Of course, in a real problem the oracle bounds are not known. Nevertheless this result motivates the use of estimators of the minimum and the maximum to estimate the conditional support so that to accelerate convergence.

7. Conclusion

In this work, we extended StoOO to a generic algorithm applicable to any functional of the reward distribution. We proposed a tailored application to the problem of quantile optimization, with four variants: one based on the classical Hoeffding's inequality, one based on Bernstein's inequality, and two others based on Chernoff's inequality. We showed that using Chernoff's inequality to build confidence intervals resulted in a dramatic improvement, both in theory and practice. We also illustrated the ability of StoROO to optimize the CVaR and compared numerically four variants.

For simplicity, we assumed that the local regularity (or at least, an upper bound) of the target function at the optimum was known to the user. However, we believe that it might

be possible to combine our results to the procedure defined in Grill et al. (2015); Xuedong et al. (2019) so as to propose an algorithm able to optimize g without the knowledge of the smoothness near an optimal point: this is left for future work. A second possible extension is to leverage the results proposed here to design an algorithm for the cumulative regret, in the spirit of HOO Bubeck et al. (2011) for example.

Acknowledgments

We would like to thank Sébastien Gerchinovitz for the discussions and his useful comments.

References

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

Peter L Bartlett, Victor Gabillon, and Michal Valko. A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. *arXiv preprint arXiv:1810.00997*, 2018.

Fabio Bellini and Elena Di Bernardino. Risk management with expectiles. *The European Journal of Finance*, 23(6):487–506, 2017.

Aharon Ben-Tal and Marc Teboulle. An Old-New Concept of Convex Risk Measures: The Optimized Certainty Equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

Clément Bouttier. Optimisation globale sous incertitudes: algorithmes stochastiques et bandits continus avec application à la planification de trajectoires d'avions. 2017.

David B Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

Yahel David and Nahum Shimkin. Pure exploration for max-quantile bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 556–571. Springer, 2016.

Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2018.

Jean-Bastien Grill, Michal Valko, and Rémi Munos. Black-box optimization of noisy functions with unknown smoothness. In *Advances in Neural Information Processing Systems*, pages 667–675, 2015.

Adam J Hepworth. *A multi-armed bandit approach to superquantile selection.* PhD thesis, Monterey, California: Naval Postgraduate School, 2017.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM, 2008.

Ravi Kumar Kolla, Krishna Jagannathan, et al. Risk-aware Multi-armed Bandits Using Conditional Value-at-Risk. *arXiv preprint arXiv:1901.00997*, 2019.

Andrea Locatelli and Alexandra Carpentier. Adaptivity to Smoothness in X-armed bandits. In *Conference on Learning Theory*, pages 1463–1492, 2018.

Rémi Munos et al. From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7 (1):1–129, 2014.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Marzena Rostek. Quantile maximization in decision theory. *The Review of Economic Studies*, 77(1):339–371, 2010.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

Balazs Szorenyi, Róbert Busa-Fekete, Paul Weng, and Eyke Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In *32nd International Conference on Machine Learning*, pages 1660–1668, 2015.

Philip Thomas and Erik Learned-Miller. Concentration Inequalities for Conditional Value at Risk. In *International Conference on Machine Learning*, pages 6225–6233, 2019.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, pages 19–27, 2013.

Shang Xuedong, Emilie Kaufmann, and Michal Valko. General parallel optimization a without metric. In *Algorithmic Learning Theory*, pages 762–787, 2019.